Data Enginering Project

Tokyo Olympic - AZURE).

Réalisé par :

Boujbair Oussamae

Supervisé par :

MOI MM

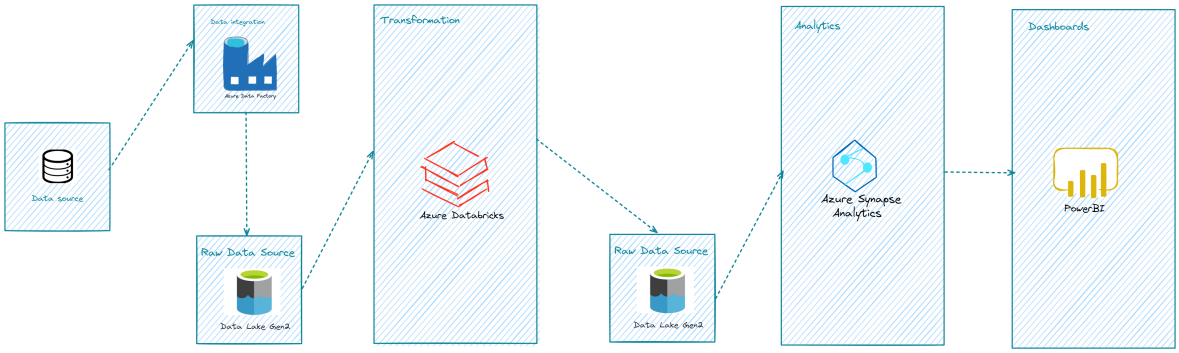
DEMOS

07 DEC 2023
Rabat, Maroc

Contents

1 Project architecture

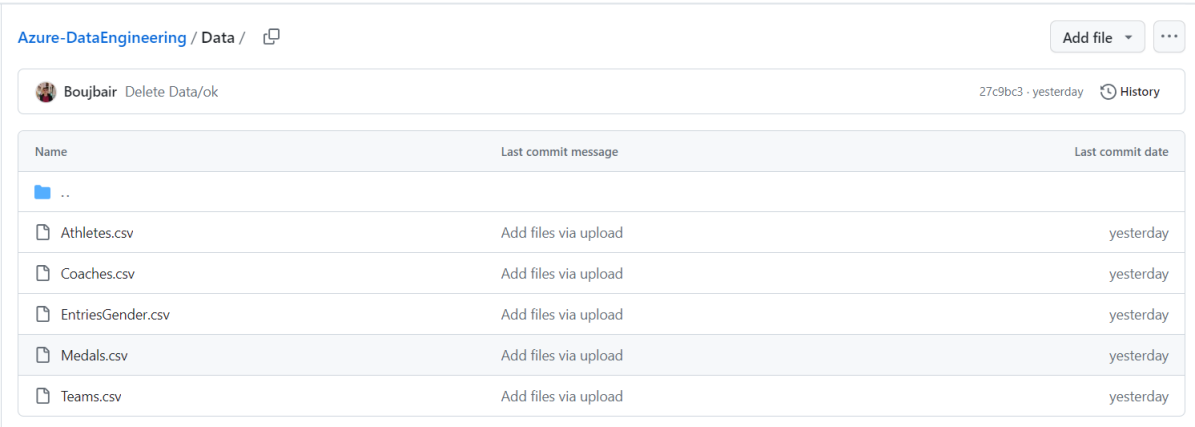
Voici l'architecture du projet:



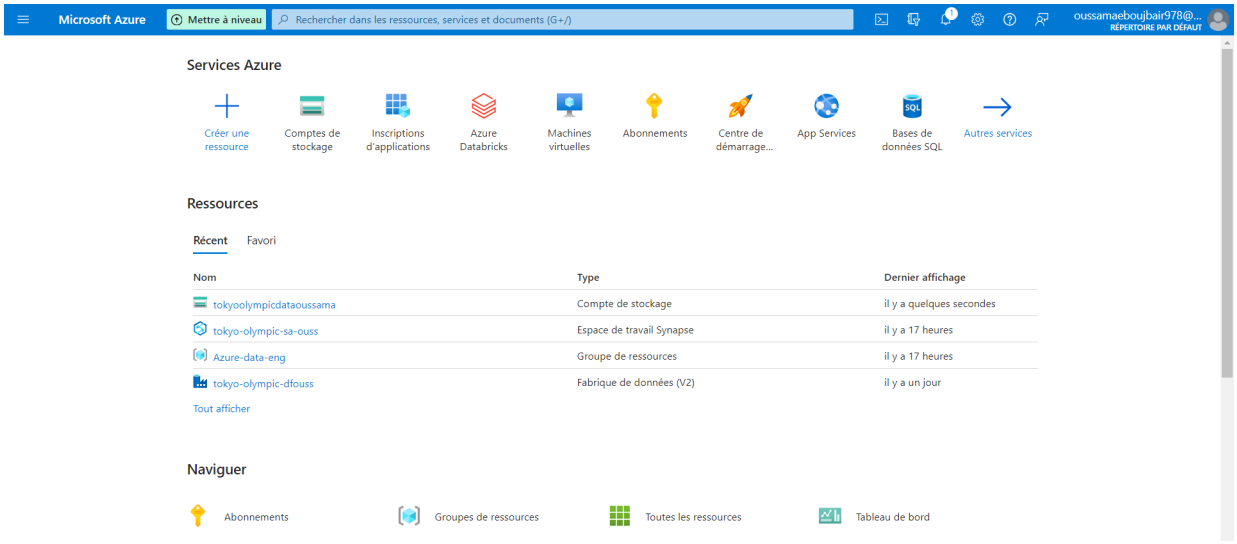
2 Data Source (GitHub Repo)

2.1 GitHub: 5 files

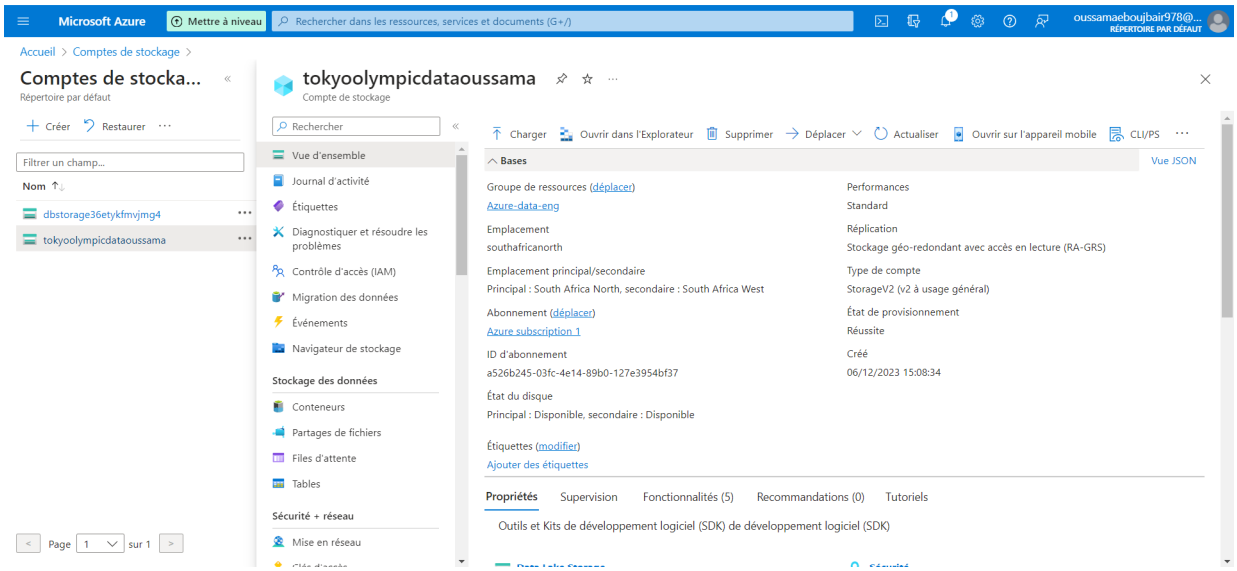
On a crée un repo sur github et on a charger 5 fichiers qu'on va exploiter par la suite:



3 Azure

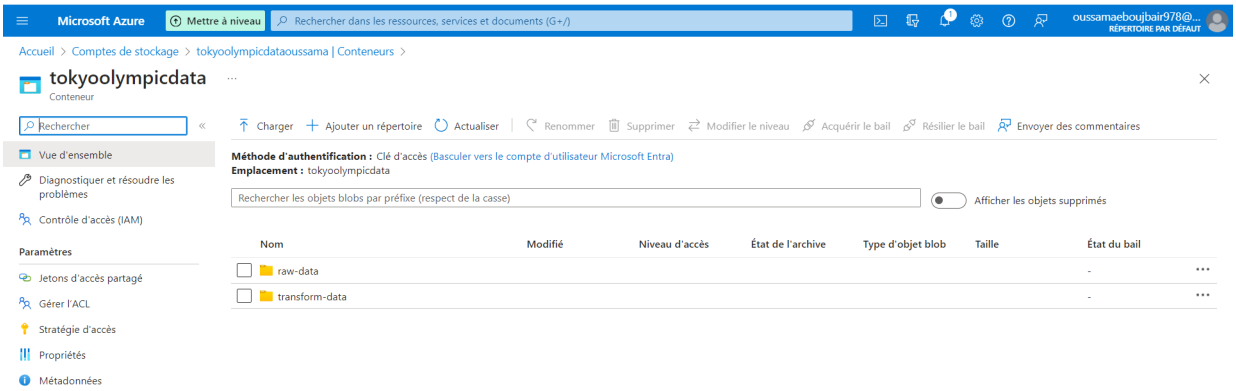


3.1 Création compte de stockage



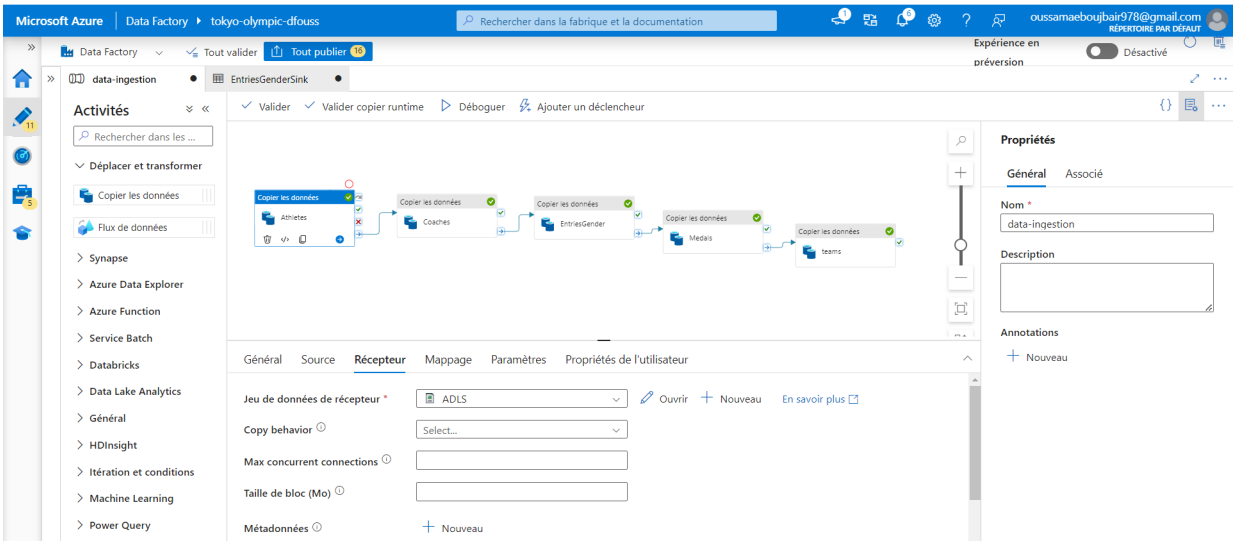
3.2 Création conteneur:

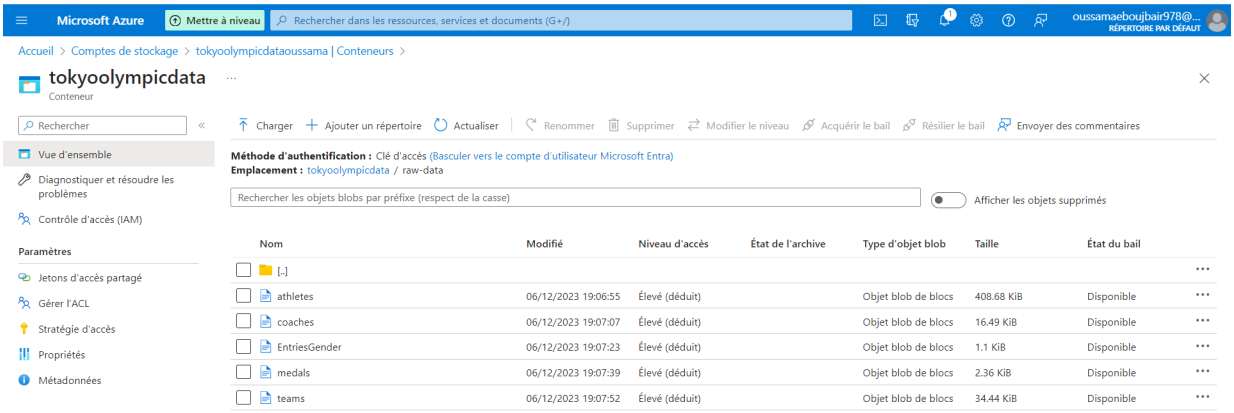
On crée un conteneur et on crée deux fichiers raw-data et transform data dans ce conteneur.



3.3 Data Factory

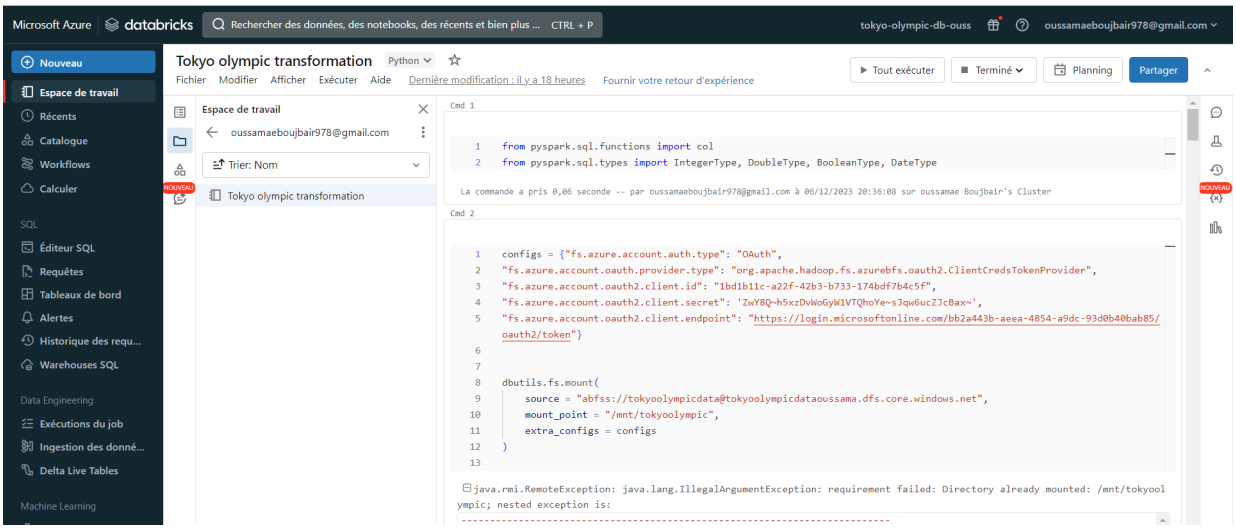
On importe les données de la source (github repo) dans data lake gen 2 (raw-data), on utilisant lien HTTP. Data factory permet d'ingester les données from diff sources, dans notre cas c'est github repo.

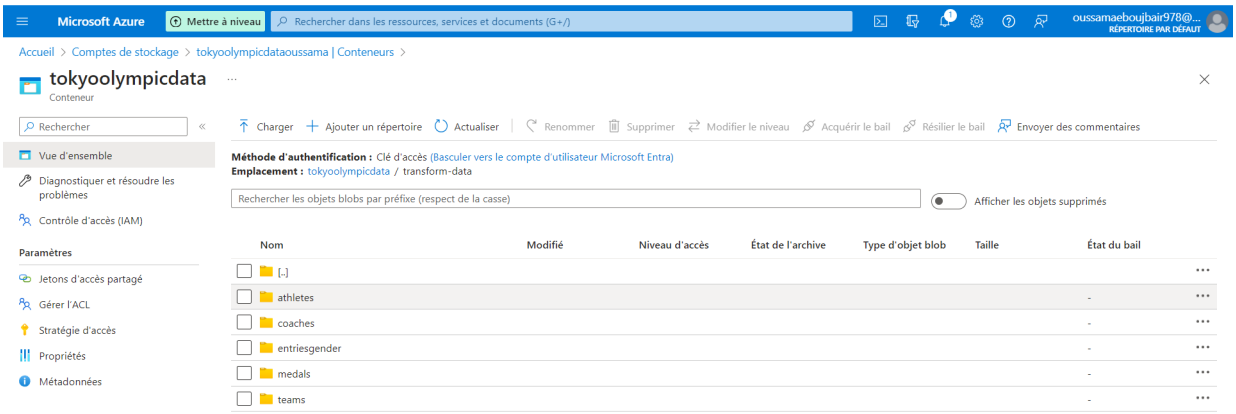




3.4 Azure Databricks

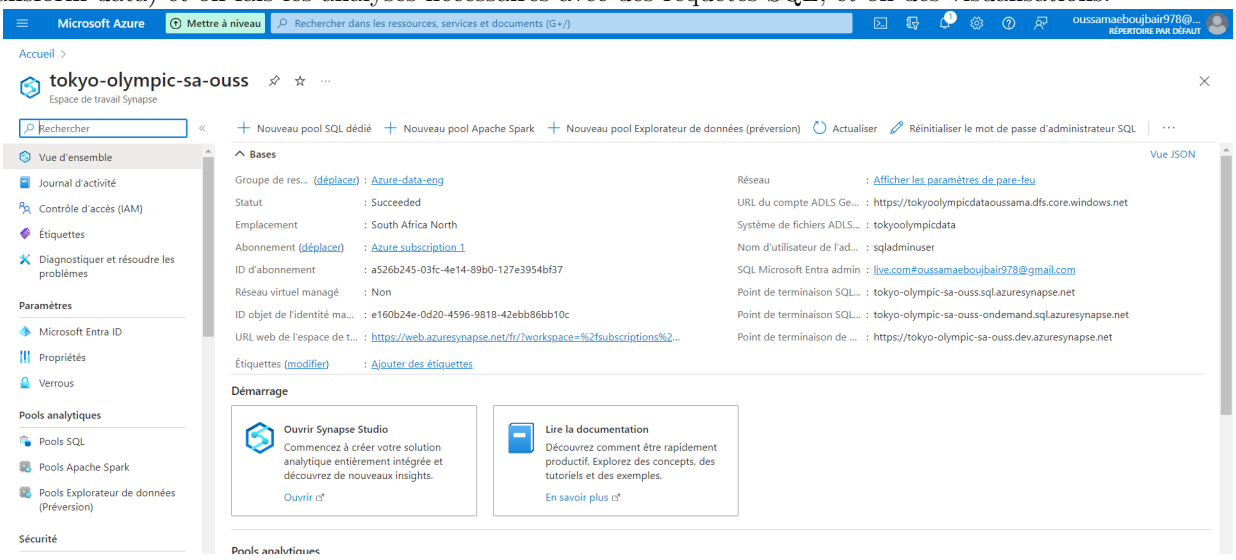
On crée un noutebook sur Databricks pour transformer les données.
On import les données de data lake gen 2 (Raw-data) avec script pyspark et on fait les transformations neccessaires et on charge les données transformées dans data lake gen2 (transform data)

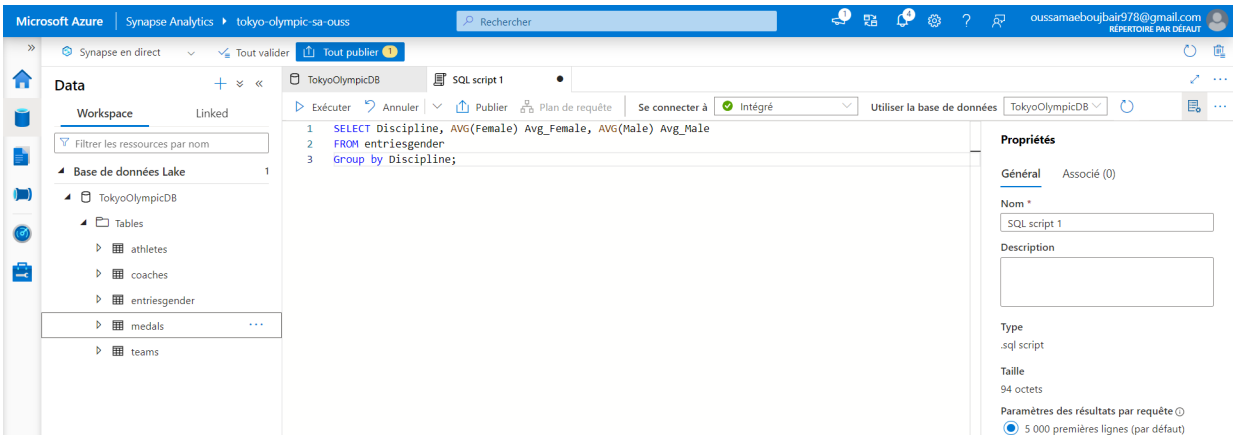


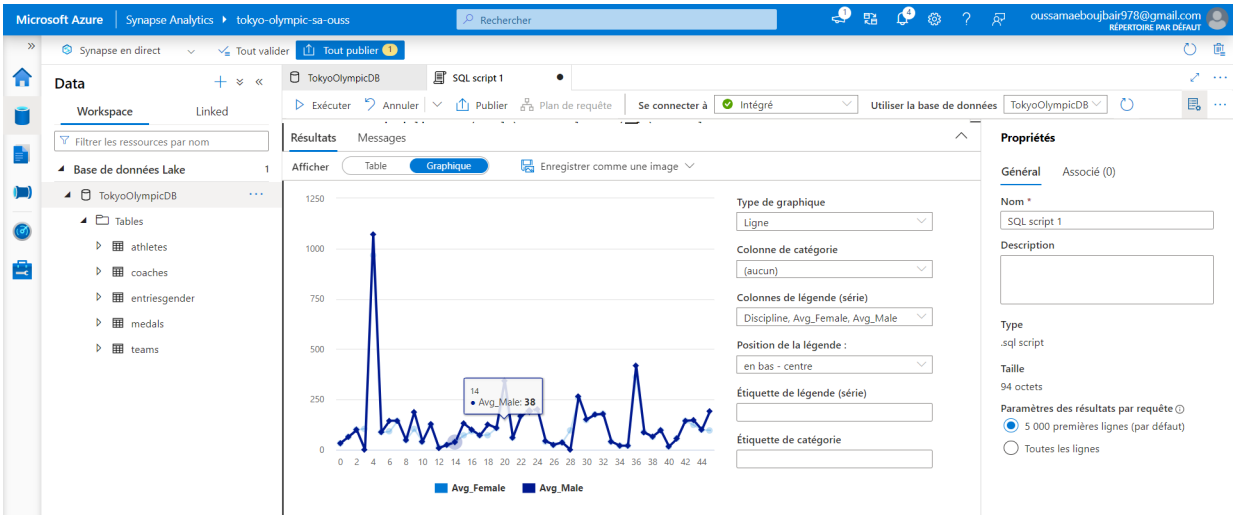


3.5 Azure synapse analytics

On crée un compte azure synapse analytics et on importe les données transformées from data lake gen 2 (transform data) et on fais les analyses necessaires avec des requetes SQL, et on des visualisations.




Microsoft Azure
Synapse Analytics
tokyo-olympic-sa-ouss
Rechercher
oussamaeboujbair978@gmail.com
RÉPERTOIRE PAR DÉFAUT
Synapse en direct
Tout valider
Tout publier
Data
Workspace
Linked
Filtrer les ressources par nom
Base de données Lake
TokyoOlympicDB
Tables
athletes
coaches
entriesgender
medals
teams
TokyoOlympicDB
SQL script 1
Exécuter
Annuler
Publier
Plan de requête
Se connecter à
Intégré
Utiliser la base de données
TokyoOlympicDB
1 SELECT Discipline, AVG(Female) Avg_Female, AVG(Male) Avg_Male
2 FROM entriesgender
3 Group by Discipline;
Propriétés
Général
Associé (0)
Nom *
SQL script 1
Description
Type
.sql script
Taille
94 octets
Paramètres des résultats par requête
5 000 premières lignes (par défaut)


Microsoft Azure
Synapse Analytics
tokyo-olympic-sa-ouss
Rechercher
oussamaeboujbair978@gmail.com
RÉPERTOIRE PAR DÉFAUT
Synapse en direct
Tout valider
Tout publier
Data
Workspace
Linked
Filtrer les ressources par nom
Base de données Lake
TokyoOlympicDB
Tables
athletes
coaches
entriesgender
medals
teams
TokyoOlympicDB
SQL script 1
Exécuter
Annuler
Publier
Plan de requête
Se connecter à
Intégré
Utiliser la base de données
TokyoOlympicDB
Résultats
Messages
Afficher
Table
Graphique
Enregistrer comme une image
1250
1000
750
500
250
0
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44
14
Avg_Male: 38
Avg_Female
Avg_Male
Type de graphique
Ligne
Colonne de catégorie
(aucun)
Colonnes de légende (série)
Discipline, Avg_Female, Avg_Male
Position de la légende :
en bas - centre
Étiquette de légende (série)
Étiquette de catégorie
Propriétés
Général
Associé (0)
Nom *
SQL script 1
Description
Type
.sql script
Taille
94 octets
Paramètres des résultats par requête
5 000 premières lignes (par défaut)
Toutes les lignes

FIN